Taylor Larrechea
Dr. Gustafson
MATH 366
CP. 19.1

1) Fixed-point system has all numbers given with a fixed number. i.e: 62.358, 0.014, 1.000
Floating-point system has some of the numbers abbreviated with powers of 10. i.e:
$0.6247 \cdot 10^{3}$, $0.1735 \cdot 10^{-13}$, $-0.2000 \cdot 10^{-1}$ etc.

2) A significant digit of a number C is any given digit of C, except possibly for zeros to the left of the first nonzero digit.

3) Example of 3D and 5S : 12.345

4)

(1) : $a = \pm m \cdot 10^{n}$, $0.1 \leq |m| < 1$, n integer.

(2) : $\bar{a} = \pm \bar{m} \cdot 2^{n}$, $\bar{m} = 0.d_1 d_2 \ldots d_K$, $d_1 > 0$.

5) The smallest positive machine number eps with $1 + eps > 1$ is called the machine accuracy.

6) Single Precision : (1.175×10^{-38} to 3.403×10^{38}). A computation of a number outside a range that occurs is called underflow when the number is smaller and overflow when it is larger.

7) An error that is caused by chopping (= discarding all digits from some decimal on) or rounding is called a roundoff error.

8) Chopping is not recommended because the corresponding error can be larger than that in rounding.

9) (3) $$\left| \frac{a - \bar{a}}{a} \right| \approx \left| \frac{m - \bar{m}}{m} \right| \leq \frac{1}{2} \cdot 10^{1-K}$$

10) Rounding errors may ruin a computation completely, even a small computation. In general, these errors become the more dangerous the more arithmetic operations (perhaps several millions!) we have to perform. It is therefore important to analyze computational programs for expected rounding errors and to find an arrangement of the computations such that the effect of rounding errors is as small as possible.

11) Accuracy in Tables. Although available software has rendered various tables of function values superfluous, some tables (of higher functions, of coefficients of integration formulas, etc.) will still remain in occasional use. If a table shows K significant digits, it is conventionally assumed that any value $\tilde{a}$ in the table deviates from the exact value a by at most $\pm \frac{1}{2}$ unit of the K^{th} digit.

12) This means that a result of a calculation has fewer correct digits than the numbers from which it was obtained. This happens if we subtract two numbers of about the same size, for example, 0.1439 - 0.1426 ("Subtractive cancellation"). It may occur in simple problems, but it can be avoided in most cases by simple changes of the algorithm - if one is aware of it! Let us illustrate this with the following basic problem.

13) Final results of computations of unknown quantities generally are approximations; that is, they are not exact but involve errors. Such an error may result from a combination of the following effects. Roundoff errors result from rounding, as discussed above. Experimental errors are errors of given data (probably arising from measurements). Truncating errors result from truncating (prematurely breaking off), for instance, if we replace a Taylor series with the sum of its first few terms. These errors depend on the computational method used and must be dealt with individually for each method. ["Truncating" is sometimes used as a term for chopping off (see before), a terminology that is not recommended.]

14) Error Bound for $\tilde{a}$, that is, a number B such that

$$|\varepsilon| \leqq B, \qquad \text{hence} \qquad |a-\tilde{a}| \leqq B$$

This tells us how far away from our computed $\tilde{a}$ the unknown a can at most lie.

15) This is an important matter. It refers to how errors at the beginning and in later steps (roundoff, for example) propagate into the computation and affect accuracy, sometimes very drastically. We state here what happens to error bounds. Namely, bounds for the error add under addition and subtraction, whereas bounds for the relative error add under multiplication and division. You do well to keep this in mind.

16) Every numeric method should be accompanied by an error estimate.

17) A stable algorithm has small changes in the initial data that cause only small changes in the final result.